

Fast-AT: Fast Automatic Thumbnail Generation using Deep Neural Networks

Seyed A. Esmaeili Bharat Singh Larry S. Davis
 University of Maryland, College Park
 {sesmaeil@umd.edu, bharat@cs.umd.edu, lsd@umiacs.umd.edu}

Abstract

Fast-AT is an automatic thumbnail generation system based on deep neural networks. It is a fully-convolutional CNN, which learns specific filters for thumbnails of different sizes and aspect ratios. During inference, the appropriate filter is selected depending on the dimensions of the target thumbnail. Unlike most previous work, Fast-AT does not utilize saliency but addresses the problem directly. In addition, it eliminates the need to conduct region search on the saliency map. The model generalizes to thumbnails of different sizes including those with extreme aspect ratios and can generate thumbnails in real time. A data set of more than 70,000 thumbnail annotations was collected to train Fast-AT. We show competitive results in comparison to existing techniques.

1. Introduction

Thumbnails are used to facilitate browsing of a collection of images, make economic use of display space, and reduce the transmission time. A thumbnail image is a smaller version of an original image that is meant to effectively portray the original image (Figure 1). Social media websites like Facebook, Twitter, Pinterest etc. have content from multiple user accounts which needs to be displayed on a fixed resolution display. A normal web page on Facebook contains hundreds of images, which are essentially thumbnails of larger images. Therefore, it is important to ensure that each thumbnail displays the most useful information present in the original image. Since images displayed on a web page vary significantly in *size* and *aspect ratio*, any thumbnail generation algorithm must be able to generate thumbnails over a range of scales and aspect ratios.

The standard operations used for creating thumbnails are cropping and scaling. Since thumbnails are ubiquitous and the manual generation of thumbnails is time consuming, significant research has been devoted for automatic thumbnail generation. Most methods [21, 4, 3] utilize a saliency map to identify regions in the image that could serve as good crops to create thumbnails. This leads to a



Figure 1. Illustration of the thumbnail problem. The original image is shown on the left with thumbnails of different aspect ratios on the right.

two step solution where saliency is first computed and then an optimization problem is solved to find the optimum crop. Whereas a recent method addresses [10] the problem directly, it involves hand crafted features and uses SVMs to score candidate crops. Moreover, the implementation requires 60 seconds to produce a single thumbnail.

We propose Fast-AT, a deep learning based approach for thumbnail generation that addresses the problem directly, in an end-to-end learning framework. Our work involves the following contributions:

- Fast-AT is based on an object detection framework, which takes dimensions of the target thumbnail into account for generating crops. Since it produces thumbnails using a feed-forward network, it can process 9 images per second on a GPU.
- By vector quantizing aspect ratios, Fast-AT learns different filters for different aspect ratios during training. During inference, the appropriate filter is selected depending on dimensions of the target thumbnail.
- For training the network, we collected a data set of 70,000 images which were cropped by workers on Amazon Mechanical Turk. We also release the annotated thumbnail dataset with this paper.

2. Related Work

Since thumbnail creation involves reducing the image size, retargeting methods [17] such as seam carving and

non-homogeneous warping can be used. However these methods produce artifacts which are often pronounced since most thumbnails are significantly smaller than the original images. Therefore, most thumbnail generation methods use a combination of cropping and scaling. Typically, automatic thumbnail generators utilize a saliency map as an indicator of important regions in the image to be cropped [21, 4, 22, 3]. Region search is then performed to find the smallest region of the image that has a total saliency above a certain threshold. A brute force approach to region search is computationally expensive, therefore approximations have been investigated such as greedy search [21], restriction of the search space to a set of fixed size rectangles [20], and binarization of the saliency map [4]. Recently, an algorithm that conducts region search in linear time has been reported [3].

However, saliency mostly ignores the semantics of the scene and does not take target thumbnail size into account. Many methods address this shortcoming through a heuristic such as selecting a crop that contains all detected faces [21] or using an algorithm that depends on both the saliency and image class. Sun et al. [22] improves the saliency map by taking the thumbnail scale into account and preserving object completeness. A scale and object aware saliency map is computed by using a contrast sensitivity function [14] and an objectiveness measure [2]; then greedy search is conducted to find the optimum region, similar to [21]. Some other approaches attempt to crop the most aesthetic part of the image [24, 15].

Huang et al. were the first to directly address the problem [10]. A data set of images and their manually generated thumbnails was collected. However, only one thumbnail size of 160×120 was considered. The solution also involved scoring a large set of candidate crops and then selecting the crop with the largest score. The implementation - albeit an unoptimized CPU code - required 60 seconds to generate a thumbnail for a single image. In addition the solution [10] is based on hand crafted features and SVM, which generally have inferior performance compared to recent deep learning based methods.

Deep convolutional neural networks have achieved impressive results on high level visual tasks such as image classification [11, 19, 9], object detection [8, 16], and semantic segmentation [12]. These architectures have not only led to significantly better results but also systems that can be deployed in real time [16]. We present a solution, based on deep fully convolutional neural networks that is learned end-to-end. We take into account varying thumbnail sizes from 32×32 to 200×200 . At test time the network can produce thumbnails at 105ms per image and shows significant improvements over the existing baselines.

3. Data Collection

We started the annotation of 28,405 images from the photo quality data set of [13] using Amazon Mechanical Turk (AMT). The set includes both high and low quality images and spans a number of categories such as humans, animal, and landscape. Target thumbnail sizes were divided into three groups- thumbnails between 32 to 64, 64 to 128, and 100 to 200, in both height and width. This leads to a range of aspect ratio from 0.5 and 2. Each image was annotated three times, with different target thumbnail sizes from each group.

The annotation was done through an interface that draws a bounding box on the original image with an aspect ratio equal to that of the thumbnail. This bounding box represents the selected crop. It is scaled down to the thumbnail size and shown to the user at the same time. The user can move the box and scale it up or down, but cannot change the aspect ratio. Restricting the bounding box (crop) to have an aspect ratio equal to the thumbnail's aspect ratio leads to more flexible annotation and avoids any possible deformation affects.

To make the interface more practical, the images were scaled down such that the height does not exceed 650 and the width does not exceed 800. The Mechanical Turk workers were shown examples of good and bad thumbnails. The examples were intended to illustrate that good thumbnails capture a significant amount of content while at the same time are easy to recognize. After the data set was collected, the thumbnail images were manually swept through and bad annotations were excluded; this led to a total of 70,049 annotations.

4. Does target thumbnail size matter?

An automatic thumbnail generating system, receives two inputs: the image and the target thumbnail. Therefore, we study the dependence between the target thumbnail and the generated crop. It is clear that the generated crop should have an aspect ratio equal to the target thumbnail's aspect ratio. Selecting a crop of a different aspect ratio could cause pronounced deformations when scaling the crop down to the thumbnail size as shown in Figure 2. It is worth noting that although deformations can be caused when selecting crops with aspect ratios different from that of the thumbnail, it has been ignored in some work [3].

Second, intuitively it is expected that smaller input thumbnail sizes would typically require smaller crops. Since larger crops would be less recognizable, the more they are scaled down. To investigate this, we compared the average area of the annotated crop vs the thumbnail size in the annotated dataset and did not find any correlation between the two. Thus, we concluded that to produce the optimal crop, the thumbnail size does not need to be taken into ac-

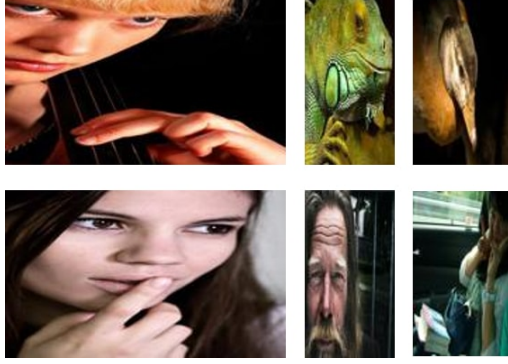


Figure 2. The above thumbnails were generated using the code from [22] which is agnostic to the thumbnail aspect ratio. The objects look clearly deformed in the thumbnails, where variation in aspect ratio is significant.

count, but the aspect ratio matters. However, we still consider a model that takes both the aspect ratio and the size of the thumbnail into account in our experiments.

5. Approach

Thumbnails are created by selecting a region in the image to be cropped (a bounding box) followed by scaling the bounding box down to the thumbnail size. We present a solution to this problem employing a deep convolutional neural network that learns the best bounding boxes to produce thumbnails. Since we formulate the problem as a bounding box prediction problem, it is closely connected to object detection. However, unlike object detection, the final predictions will not consist of bounding boxes with a discrete probability distribution across the different classes, but involve two classes: representative of the image vs non-representative of the image.

Early deep learning methods for object detection utilized proposal methods such as [23] that were time-consuming. Computation time was significantly reduced using region proposal networks (RPNs) [16] that learn to generate proposals. In addition, the computational expense of forward propagating the pooled proposal features through two fully connected layers has been reduced by introducing a new convolutional layer consisting of class-specific position-sensitive filters [5]. Specifically, if there are C classes to be detected, then this new convolutional layer will produce $k^2(C + 1)$ feature maps. The k^2 position-sensitive score maps correspond to $k \times k$ evenly partitioned cells of objects. Those k^2 feature maps are associated with different relative positions in the spatial grid, such as (top-left,...,bottom-left) for every category. $k = 3$, corresponds to a 3×3 spatial grid and 9 position-sensitive filters per class. Every class (including the background), will have k^2 feature maps associated with it. Instead of forward propagating through

two fully connected layers, positive-sensitive pooling followed by score averaging is performed, a $(C + 1)$ -d vector is generated and used to predict softmax responses across the categories.

An architecture for thumbnail generation should be fully convolutional because, including a fully connected layer requires a fixed input size. If there were a mismatch between the aspect ratio of an image and the fixed input size, the image would have to be cropped in addition to being scaled. Because the thumbnail crops (bounding boxes) could touch the boundaries of the image and even extend to the whole image, the pre-processing step of cropping a region of the image could lead to sub-optimal predictions, since part of the image has been removed. Therefore, an architecture similar to [19] that was used for the image net localization challenge [6], which simply replaces the class scores by 4-D bounding predictions cannot be employed, because of the fully connected layer at the end.

Another observation is that unlike object detection, the thumbnail generation network receives two inputs: the image and the thumbnail aspect ratio. Both RPN and R-FCN introduce task-specific filters. In the case of RPN, filter banks that specialize in predicting proposals of a specific scale are achieved by modifying the training policy. In the case of R-FCN, position-sensitive filters specialize through the position-sensitive pooling mechanism. In a similar manner, we modify R-FCN for thumbnail creation by introducing a set of aspect ratio-specific filter banks. A set of A points are introduced in the aspect ratio range of $[0.5, 2]$, which represent aspect ratios that grow by a constant factor (a geometric sequence), i.e. it is of the form $S = \{\frac{1}{2}c, \frac{1}{2}c^2, \dots, \frac{1}{2}c^A\}$. Note that $\frac{1}{2}c^0 = \frac{1}{2}$ and $\frac{1}{2}c^{(A+1)} = 2$, leads to $c = \sqrt[A+1]{4}$. The filter banks in the last convolutional layer in R-FCN are modified into A pairs, with each pair having a total of k^2 filters. Each pair is associated with a single element in the set S . Similar to R-FCN, position-sensitive pooling followed by averaging is performed over that pair and those two values are used to yield a softmax prediction of representativeness.

At training time, when an image, thumbnail size pair is received, the image is forward propagated through convolutional layers up to the last convolutional layer. The thumbnail aspect ratio is calculated and the element with the closest value is selected from S - the pair associated with that element factors in the training while others are ignored. For this pair, the proposals are received, and similar to object detection, positive and negative labels are assigned to the proposals based on their intersection over union (IoU) with the ground truth. Specifically, a positive label is assigned if the $\text{IoU} \geq 0.5$ and negative otherwise. Similarly, A aspect ratio-specific regressors are trained, one for each element in S ; these are similar to class-specific regressors. For a given

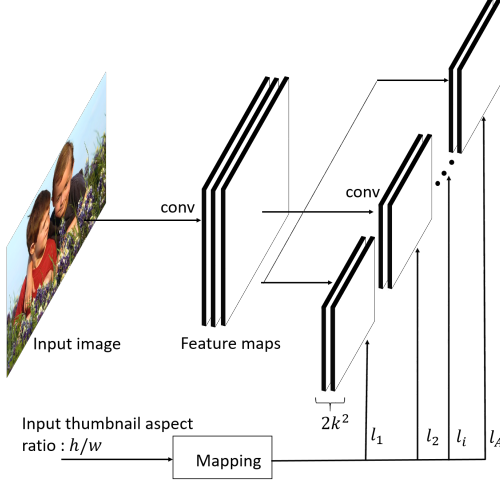


Figure 3. Illustration of the Fast-AT architecture and training policy. The appropriate filter is decided based on the thumbnail aspect ratio.

proposal, we employ the following loss:

$$L(s_i, t_i) = \sum_{i=1}^A l_i L_{cls}(s_i, s^*) + \lambda [s_i^* = 1 \wedge l_i] L_{reg}(t_i, t^*)$$

, where l_i is either ignore=0 or factor-in=1, namely

$$l_i = \begin{cases} 1 & \text{if } i = \operatorname{argmin}_i |\frac{1}{2}c^i - \text{thumbnail aspect ratio}| \\ 0 & \text{otherwise} \end{cases}$$

s_i is the representativeness score predicted by the i^{th} pair, s^* is the ground truth label, and L_{cls} is cross entropy loss. λ is a weight for the regression loss which we set to 1. The regression loss is 0 for all but the nearest aspect ratio. For the filter corresponding to the nearest aspect ratio, L_{reg} is the smooth L_1 loss as defined in [7]. t_i is the bounding box predictions made by the i^{th} regressor and t^* is the ground truth bounding box. Both predictions are parametrized as in [7]. Figure 3 illustrates the architecture.

Since each regressor is responsible for a range of input thumbnail sizes, the predictions made by any regressor at test time could have an aspect ratio that differs from the target thumbnail aspect ratio. Therefore the output bounding box has to be rectified to have an aspect ratio equal to the thumbnail’s aspect ratio, to eliminate any possible deformations when scaling down. We employ a simple method where a new bounding box with an aspect ratio equal to that of the target thumbnail is placed at the center of the predicted box and is expanded until it touches the boundaries. Since the predicted box already has an aspect ratio close to that of the thumbnail, the difference between the rectified box and the predicted box is not significant, as shown in Figure 4(b).

Our implementation of Fast-AT is based on Resnet-101 [9], a learning rate of 0.001, momentum of 0.9, weight decay of 0.0005 with approximate joint training is used [16].

5.1. Does R-FCN alone work?

Among the baselines we consider is R-FCN- without any modifications. In effect, it is performing object detection between two classes. We find that R-FCN alone generates bounding boxes that have good representations of the original image. But since the architecture is agnostic to the input thumbnail dimension, the generated thumbnails are of low quality as shown in Figure 4(a). If we apply the same rectification to the generated boxes, to cancel the deformation affects, important parts of the images are not preserved, in contrast to our model’s results which are shown in shown in Figure 4(b). This is because of the significant mismatch between the target thumbnail aspect ratio and the predicted box aspect ratio. Eliminating the rectification step would lead to deformed results, as shown in Figure 2.

6. Experiments

In comparing models we use the following metrics:

- offset: the distance between the center of the ground truth bounding box and the predicted bounding box.
- rescaling factor (rescaling): defined as $\max(s_g/s_p, s_p/s_g)$ where s_g and s_p are the rescaling factors for the ground truth and predicted box, respectively. [10].
- IoU: intersection over union between the predicted box and the ground truth.
- aspect ratio mismatch (mismatch): the square of the difference between the aspect ratio of the predicted box and the aspect ratio of the thumbnail.

Comparisons between the models is shown in Table 1. The first model we use is R-FCN without modification. This architecture is agnostic to thumbnail dimension, the number of classes are reduced to two and the architecture is modified accordingly. We see that R-FCN alone has good performance in terms of all metrics except for aspect ratio mismatch. High values in these metrics show that rectifying the bounding box will cause significant change in the predicted box. We next consider our proposed model, where we map based on aspect ratio, with 5 divisions ($A = 5$). We see significant improvements in the metrics - the mean IoU has increased by 4% and the offset and rescaling factor have been reduced. The aspect ratio mismatch has also been significantly reduced. We extend the divisions to thumbnail sizes as well. In this case we divide the input thumbnail space into three branches: small thumbnails (32-64),

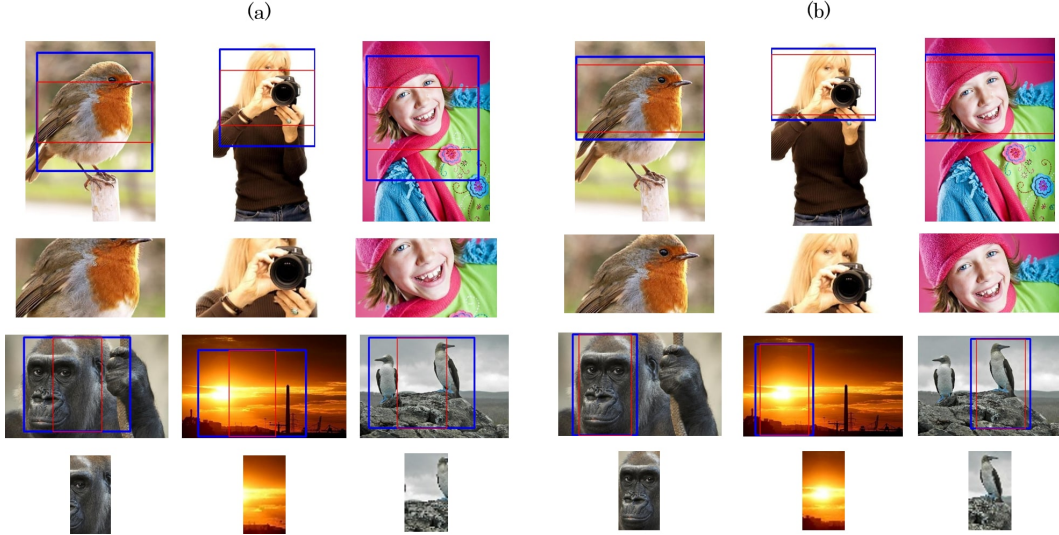


Figure 4. R-FCN and Fast-AT predictions. (a):The original image is displayed with the R-FCN prediction in blue, the rectified box is in red and the resulting thumbnail is below. Note how the resulting thumbnail is missing important parts of the original image.(b):The original image is displayed with the Fast-AT prediction in blue, the rectified box is in red and the resulting thumbnail is below. The rectification does not introduce a significant change in the box predicted by Fast-AT.

| Model | offset | rescaling | IoU | mismatch |
|-------------------------------|--------|-----------|------|----------|
| R-FCN | 56.2 | 1.192 | 0.64 | 0.102 |
| Fast-AT (AR) | 55.0 | 1.149 | 0.68 | 0.010 |
| Fast-AT (AR+TS) | 55.4 | 1.154 | 0.68 | 0.012 |
| Fast-AT (AR, scale up of 350) | 53.1 | 1.156 | 0.69 | 0.024 |

Table 1. Metrics computed using different models. R-FCN, Fast-AT with aspect ratio mapping (AR), Fast-AT with aspect ratio and thumbnail size mapping (AR+TS), and Fast-AT with aspect ratio mapping scaled up to 350.

medium (64-100), and large (100-200). Each branch is further divided based on aspect ratio as in the first model. This leads to a total of $5 \times 3 = 15$ regressors and 15 pairs of $2k^2$ filters. This did not lead to an improvement over the model with only aspect ratio divisions.

Unlike object detection bounding boxes, the predicted bounding boxes for thumbnails can enclose multiple objects and may extend to the whole image. Therefore, a network with a small receptive field may predict accurate bounding boxes for object detection, but its predictions for thumbnail crops may be inaccurate. For object detection, Faster RCNN [16] effectively reduces the receptive field of the network by scaling up the image so that the smallest dimension is 600. This step is further implemented in R-FCN as well. We reduce the image dimension (minimum of height/width) from 600 to 350 in Fast-AT to investigate the affect of the receptive field. We observe a slight improvement in the offset and IoU as shown in Table 1. The improvement is not large because we use Resnet-101 [9], which already has a large receptive field. Fast-AT’s architecture is generic and hence other models such as VGGNet [19] or ZFNet [25] can also be used. Such shallower models are likely to ben-

efit significantly in thumbnail generation if their receptive fields are extended.

| Model | offset | rescaling | IoU | mismatch |
|-----------------|--------|-----------|------|----------|
| Fast-AT (AR) | 55.9 | 1.149 | 0.67 | 0.011 |
| Fast-AT (AR+TS) | 55.0 | 1.153 | 0.68 | 0.012 |

Table 2. Performance Comparison between Fast-AT with aspect ratio mapping (AR) and Fast-AT with aspect ratio and thumbnail size mapping (AR+TS) at small thumbnail sizes below 64×64 .

We further compare Fast-AT with aspect ratio divisions and Fast-AT with aspect ratio and thumbnail divisions over thumbnails with a small size; below 64×64 . We do not see a significant improvement (Table 2). This further confirms our initial conclusion that it is the thumbnail aspect ratio not the the thumbnail size that matters.

We also measure the metrics at extreme aspect ratios, i.e. aspect ratios below 0.7 and above 1.8, the results are shown in Table 3. We observe a significant drop in R-FCN’s performance, namely, IoU falls by about 6% and the mismatch almost doubles. At the same time Fast-AT still performs well. This shows that Fast-AT can handle thumbnails of

widely varying aspect ratios.

| Model | offset | rescaling | IoU | mismatch |
|-----------------|--------|-----------|------|----------|
| R-FCN | 57.5 | 1.348 | 0.58 | 0.200 |
| Fast-AT (AR) | 49.8 | 1.18 | 0.68 | 0.013 |
| Fast-AT (AR+TS) | 50.7 | 1.183 | 0.68 | 0.014 |

Table 3. Performance Comparison between R-FCN, Fast-AT with aspect ratio mapping (AR) and Fast-AT with aspect ratio and thumbnail mapping (AR+TS) at extreme values of aspect ratio (above 1.8 and below 0.7).

7. Evaluation

We compare our method to the other methods, by metric evaluations and by visual results. We compare against 3 methods:

- **Direct-Downsizing:** In this method a crop with an aspect ratio equal to the thumbnail is placed at the center and expanded until it touches the image boundaries.
- **Scale and Object-Aware Saliency (SOAT):** In this method scale and object-aware saliency is computed and a greedy algorithm is used to conduct region search over the generated saliency map [22].
- **Efficient Cropping:** This method generates a saliency map and conducts region search in linear time. Unlike previous methods, the search can be restricted to regions with a specific aspect ratio. However, when the aspect ratio mismatch between the image and the thumbnail is significant, a solution may not exist. In such cases, we apply the method without aspect ratio restriction. We run this method with three different values of saliency threshold: 0.7, 0.8, and 0.9.

We could not compare against [10] and [24] because the code was not released.

7.1. Metrics

For comparing the different methods, we use the same metrics that were used in the experiment. In addition, we use the hit ratio h_r and the background ratio b_r [10], which are defined as:

$$h_r = \frac{|g \cap p|}{|g|} \quad \text{and} \quad b_r = \frac{|p| - |g \cap p|}{|g|}$$

where g is the ground truth box and p is the predicted box. The metrics are computed over an annotated test set consisting of 7,005 images and averaged. These images were not used in training Fast-AT. Table 4 shows the performance of different methods. Note that offset is higher than that reported in [10]. Unlike the MIRFLICKR-25000 dataset [1] that was used in [10], the data set we use has images

with larger variation in size, some with low quality, and it includes many images with multiple objects. In addition, our thumbnails have an aspect ratio that varies from 0.5 to 2. This makes the data set significantly more challenging and would explain the large increase in the offset values compared to the reported results in [10]. We find that our method performs the best in terms of offset, rescaling factor, and IoU. Because direct-downsizing predicts a bounding with an aspect ratio equal to the target, it is the only method which outperforms Fast-AT in terms of aspect ratio mismatch. We note that efficient cropping has a non-zero aspect ratio mismatch, indicating that there were examples where the problem was infeasible when the aspect ratio restriction is imposed. This is expected given the wide variation of thumbnail aspect ratios in our data set. Efficient cropping performs better in terms of offset and intersection over union with higher threshold values. Note that higher threshold values are more likely to lead to an infeasible problem which leads to the observed increase in aspect ratio mismatch. Unsurprisingly, SOAT which is agnostic to the target aspect ratio has the highest aspect ratio mismatch.

The hit ratio represents the percentage of ground truth that was captured by the bounding box and the background ratio represents the percentage of bounding box area that lies outside the ground truth. The optimal method should be close to the ground truth and therefore should have a large hit ratio and a small background ratio. We find that the performance of different methods in terms of hit and background ratios is similar to the results reported in [10]. Namely, the direct-downsizing method has the highest ratios which is explained by its very large bounding box predictions that are restricted only by the aspect ratio mismatch between the image and the target thumbnail. The saliency based methods focus on a relatively small region having large saliency. This leads to small crops which explain the low values for the hit and background ratios. Our method, in comparison is distinguished by a large hit ratio and a low background ratio. This shows that it closely matches the predicted ground truth boxes.

We observe that the direct-downsizing method outperforms the saliency based methods (SOAT and efficient cropping) in terms of the metrics, suggesting that saliency based cropping can miss representative regions in the images while emphasizing recognizability over representation. The saliency algorithms are not state-of-the-art used in the provided code, which may also be a contributing factor.

7.2. Visual Results

We show qualitative results in comparison to other baselines in Figure 5(1-12). The direct-downsizing method is completely agnostic to image content. This limitation leads to poor performance, especially when important content exists at image boundaries. Note how direct-downsizing

| Method | offset | rescaling factor | IoU | aspect ratio mismatch | h_r | b_r |
|------------------------------------|--------|------------------|------|-----------------------|-------|-------|
| Direct-downsizing | 67 | 1.188 | 0.63 | 0 | 88.4% | 60.3% |
| SOAT | 80.5 | 1.378 | 0.52 | 0.204 | 68.7% | 41.6% |
| Efficient Cropping (threshold=0.7) | 88.3 | 1.329 | 0.52 | 0.176 | 64.4% | 34.3% |
| Efficient Cropping (threshold=0.8) | 80.2 | 1.354 | 0.56 | 0.224 | 74.2% | 45.8% |
| Efficient Cropping (threshold=0.9) | 72.2 | 1.417 | 0.59 | 0.231 | 84.7% | 63.2% |
| Fast-AT (aspect ratio) | 55.0 | 1.148 | 0.68 | 0.010 | 83.7% | 37.1% |

Table 4. Metrics evaluated on different thumbnail generation methods.

thumbnails crop out important parts of the image in the first three examples. In the first example, the face was completely cropped out. In other examples, its performance is adequate.

We see that saliency based methods improve over direct-downsizing at times. Efficient cropping has a better thumbnail for the third example. Though these methods succeed in preserving important content, in some examples however, their final thumbnails can have pronounced deformations. This can be seen in many examples for SOAT and in some examples for efficient cropping. In addition, these methods ignore the semantics of the scene and may ignore important parts of the image. This can be seen in the third and the sixth examples for SOAT and in the first and fourth examples for efficient cropping.

It can be seen from the examples that Fast-AT succeeds in each case. It preserves the content of the scene and predicts thumbnails that tightly enclose the most representative part of the image. We do not find an example where the prediction misses important regions in the image or causes the output thumbnail to be deformed. A particularly interesting example is Figure 5(10), where the target thumbnail has a small aspect ratio (a tall thumbnail), however the spread of information in the image is horizontal. This makes this example more difficult for generating a good thumbnail. However, Fast-AT still makes a better prediction than the prediction made by the direct-downsizing method, SOAT, or efficient cropping that produce a thumbnail which either misses important regions in the image or is clearly deformed.

8. Failure Cases and Multiple Predictions

We also investigate failure cases of Fast-AT. We look for examples in the test set where the prediction has an IoU with the ground truth below 0.1. Figure 5(13:a) shows some examples. The ground truth box is in green and the prediction is in blue. We see that although the prediction is very different from the ground truth, in some cases it still predicts crops that capture representative regions in the original images. Furthermore, for some of these failure cases we take the prediction with the second or third highest confidence. Figure 5(13:b) shows examples where the second or third predictions are close to the ground truth. This suggests that if the system is to be deployed, users could benefit if the sys-

tem outputs a set of top predictions instead of one. These predictions can be treated as a set of candidates which the user picks the best solution from.

Another interesting case is when there is a significant aspect ratio mismatch between the representative part of the image and the thumbnail’s aspect ratio. Because of the significant aspect ratio mismatch, the crop cannot capture all of the representative part of the image. We show that our algorithm is capable of producing multiple crops that cover different representative parts of the image. Figure 5(14) shows some examples. In the first three images (from the left), the region of interest is spread horizontally, but the thumbnail’s aspect ratio is very small (tall thumbnail). The reverse is true for the rightmost image. The first row shows the bounding box prediction with the highest confidence and the second row shows the bounding box prediction with the second highest confidence. We see that these predictions cover different representative parts of the image.

9. Future Work

Yan et al [24] considered cropping the image to improve the aesthetic quality of the image and Santella et al [18] considered cropping for preserving only the important regions in the image. These two problems have a strong connection to automatic thumbnail generation. We think it is possible to design one architecture that is capable of generating crops for different tasks. This can be done by extending our idea of having specialized aspect ratio filters to filters that specialize for different tasks as well. This investigation would require a large data set that is specific to both tasks.

10. Conclusion

In this paper we have improved over existing baselines for automatic thumbnail generation. We presented a solution that does not depend on saliency or heuristic considerations but rather attacks the problem directly. A large data set consisting of 70,049 annotations was collected. A CNN capable of generating thumbnails in real time was trained using this set. Metric and qualitative evaluations have shown superior performance over the existing methods. We have further investigated with failure examples of our model and have seen that in some failure cases the second and third predictions can be close to the ground truth.

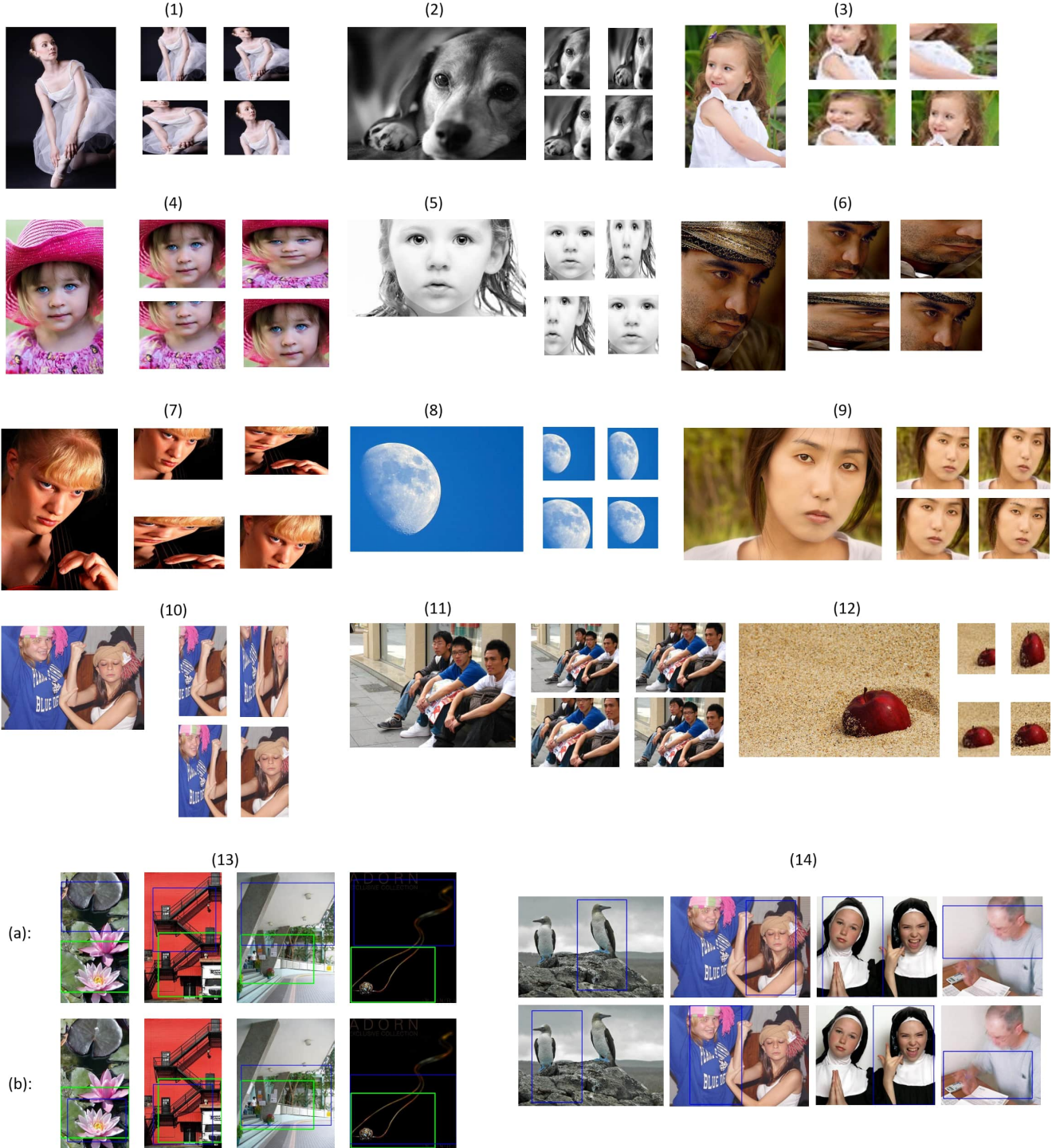


Figure 5. (1-12): Images and their generated thumbnails: The original image is on the left, to its right we display the thumbnails: top left is direct-downsizing, top right is SOAT, bottom left is efficient cropping, and bottom right is Fast-AT. (13): Failure cases of Fast-AT: (a): the ground truth is in green and the prediction is in blue. (b): we show that the second or third most confident prediction is close to the ground truth. (14): In these images there is a significant aspect ratio mismatch between the region of interest in the image and the thumbnail’s aspect ratio. The predictions are successful in covering different representative regions in the image.

References

- [1] The mirflickr retrieval evaluation. <http://press.liacs.nl/mirflickr/>. 6
- [2] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2189–2202, 2012. 2
- [3] J. Chen, G. Bai, S. Liang, and Z. Li. Automatic image cropping: A computational complexity study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 507–515, 2016. 1, 2
- [4] G. Ciocca, C. Cusano, F. Gasparini, and R. Schettini. Self-adaptive image cropping for small displays. *IEEE Transactions on Consumer Electronics*, 53(4):1622–1627, 2007. 1, 2
- [5] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. *arXiv preprint arXiv:1605.06409*, 2016. 3
- [6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 3
- [7] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 4
- [8] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2
- [9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 2, 4, 5
- [10] J. Huang, H. Chen, B. Wang, and S. Lin. Automatic thumbnail generation based on visual representativeness and foreground recognizability. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 253–261, 2015. 1, 2, 4, 6
- [11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2
- [12] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 2
- [13] W. Luo, X. Wang, and X. Tang. Content-based photo quality assessment. In *2011 International Conference on Computer Vision*, pages 2206–2213. IEEE, 2011. 2
- [14] J. Mannos and D. Sakrison. The effects of a visual fidelity criterion of the encoding of images. *IEEE Transactions on information theory*, 20(4):525–536, 1974. 2
- [15] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato. Sensation-based photo cropping. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 669–672. ACM, 2009. 2
- [16] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2, 3, 4, 5
- [17] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir. A comparative study of image retargeting. In *ACM transactions on graphics (TOG)*, volume 29, page 160. ACM, 2010. 1
- [18] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen. Gaze-based interaction for semi-automatic photo cropping. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 771–780. ACM, 2006. 7
- [19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 3, 5
- [20] F. Stentiford. Attention based auto image cropping. In *Workshop on Computational Attention and Applications, ICVS*, volume 1. Citeseer, 2007. 2
- [21] B. Suh, H. Ling, B. B. Bederson, and D. W. Jacobs. Automatic thumbnail cropping and its effectiveness. In *Proceedings of the 16th annual ACM symposium on User interface software and technology*, pages 95–104. ACM, 2003. 1, 2
- [22] J. Sun and H. Ling. Scale and object aware image thumbnailing. *International journal of computer vision*, 104(2):135–153, 2013. 2, 3, 6
- [23] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 3
- [24] J. Yan, S. Lin, S. Bing Kang, and X. Tang. Learning the change for automatic image cropping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 971–978, 2013. 2, 6, 7
- [25] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014. 5